The Report Committee for Nadine Suzanne Gibson Certifies that this is the approved version of the following Report:

Regression Analysis of Spatially Autocorrelated Data: A Study of County-Level Turnout in Texas

APPROVED BY

SUPERVISING COMMITTEE:

Tse-Min Lin, Supervisor

Robert Luskin

Regression Analysis of Spatially Autocorrelated Data: A Study of County-Level Turnout in Texas

by

Nadine Suzanne Gibson

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN STATISTICS

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2018

To Mom, Dad, and Darla.

Acknowledgments

I would like to thank Prof. Tse-Min Lin for his perseverance and commitment to helping me finish this Master's program. Without his encouragement, I would never have thought myself capable of taking on this endeavor. I would also like to thank Prof. Robert Luskin for always encouraging me to continue my methodological training even after major setbacks. I would like to acknowledge Prof. Daron Shaw for his guidance throughout my graduate career. Finally, without Rachel German, Jaclyn Peek, and Robert Shaffer I have no doubt that this would have been impossible.

Regression Analysis of Spatially Autocorrelated Data: A Study of County-Level Turnout in Texas

Nadine Suzanne Gibson, M.S.Stat.
The University of Texas at Austin, 2018

Supervisor: Tse-Min Lin

Interest in spatially weighted regression analysis has increased due to corresponding increases in access to publicly available spatial data. Spatial autocorrelation occurs when the ordering of observations across space produces a relationship between pairs of individual observations. Instances of spatial autocorrelation necessitate the use of alternative approaches to parameter estimation other than ordinary least squares. With a focus on autocorrelation resulting from spatial dependence in the dependent variable or the error term, this report summarizes basic methodology for detecting spatial autocorrelation and spatial autoregressive model selection. The approaches outlined in this report are then applied to an analysis of county-level turnout in Texas.

Table of Contents

List of Tables

List of Figures

Chapter 1

Spatial Autocorrelation in Regression Analysis

With dramatic increases in the amount of publicly available geographic information system (GIS) data and software, it has become more important for researchers to understand the complexities of analyzing spatial data. Since the advent of the internet, the average individual has access to an ever increasing supply of high quality spatial data. More recently, through the use of smartphone technology, GIS data has become an integral part of everyday life. Spatial data is not, however, new in statistical research. Data sources such as the U.S. Census and election results have been produced since the founding of the United States.[1]

The term spatial refers to how areal units are arranged on a planar map (Griffith 1987, p. 10). Autocorrelation occurs when the ordering of observations produces a relationship between pairs of individual observations. Formally, autocorrelation means

$$h_i = f(h_j), i \neq j \tag{1.1}$$

[1]What we know today as the "Census" was originally mandated by the U.S. Constitution. Article I, Section 2 of the United States Constitution, states: "The actual enumeration shall be made within three years after the first meeting of the Congress of the United States, and within every subsequent term of 10 years, in such manner as they shall by Law direct."

where an individual observation h_i is a function of other observations.

In political science research, especially research involving elections, there is important geographic variation. With more attention being paid to election administration since the 2000 election, there is an increasing need for political scientists to be aware of how to account for geographic dependencies in the regression analysis. Controlling for spatial autocorrelation is also common practice in quantitative geographic, medical, and demographic research.[2]

This report will proceed first by defining and modeling spatial autocorrelation in regression analysis. The focus of this report will be centered on spatial autoregressive processes. In particular, autocorrelation resulting from spatial dependence in the dependent variable and the disturbances. Second, this report will outline basic methods for detecting spatial autocorrelation. Third, this report will describe how the Lagrange Multiplier test can be used for autoregressive model selection. Last, this report will implement these methods in a case study of turnout in the state of Texas.

1.1 Spatial Dependence in the Regression Model

When estimating regression coefficients, in order for the OLS estimator to be considered the best linear unbiased estimator (BLUE), certain conditions

[2]Concerns over spatial autocorrelation is particularly evident in research on communicable diseases in medical statistics.

must be met. Consider the linear regression model:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1.2}$$

such that the dependent variable, y_i is a linear function of x_i with some error, ϵ_i. The Guass-Markov Theorem states that, under the following conditions,

1. Zero mean: $E(\varepsilon_i) = 0$
2. Nonstochastic X: values of X_i are fixed in repeated sampling
3. Homoskedasticity: $E(\varepsilon_i^2) = \sigma^2$
4. Non-autocorrelation: $Cov(\varepsilon_i, \varepsilon_j) = 0$, $(i \neq j)$

the OLS estimator of β is BLUE.[3]

When working with data that can be spatially mapped, it is important to test for violations of the fourth assumption (non-autocorrelation). If the fourth assumption is violated, i.e. if $Cov(\varepsilon_i, \varepsilon_j) \neq 0$, where $i \neq j$, then autocorrelation is present. If i and j are observations in time, then the autocorrelation is temporal. If i and j are observations in space, then the autocorrelation is spatial. As cautioned by Berry (1993), spatial autocorrelation "should be suspected whenever the positions of observations under analysis are structured relative to one another in the same manner". When observations are autocorrelated, the coefficient estimates remain unbiased, but are no longer efficient.

[3] The classical linear regression assumptions as articulated by Kmenta (1971) includes the additional assumption of Normality: $\varepsilon_i \sim N(0, \sigma^2)$.

1.2 Spatial Dependence

There are two types of spatial autocorrelation that researchers should investigate when using spatial data. The first type of spatial autocorrelation is in the dependent variable. The second type is in the regression error term. Substantively, the difference between these two types of spatial autocorrelation relates to the functional form of the spatial processes (Griffith 1987).

When spatial autocorrelation is observed in the dependent variable, it is because the data is organized in such a way that observations' placement and proximity to one another are non-random. Equation 1.3 presents the functional form of autocorrelation in the dependent variable,

$$\begin{aligned} y_i &= f(y_1, y_2, ..y_n), i \notin N \\ N &= \{1, 2, .., n\} \end{aligned} \tag{1.3}$$

where y_i is a function of the values of other observations of the random variable Y at other locations, $y_1, y_2, ..y_n$. When this is the case, there is clustering of similar (positive spatial autocorrelation) or dissimilar (negative spatial autocorrelation) observations.

Odland (1988, p. 53) defines spatial autocorrelation in the error term as instances where "the error at each location depends on the errors at other locations". This generally occurs when the spatial process generating autocorrelation is caused by some unobserved variable. Consider the linear regression model in Equation 1.2 where ϵ_i is correlated with ϵ_j and $i \neq j$. When the errors in one point in space, i, are dependent on another location, j, the errors of the regression model are no longer independent. The model in Equation

1.2 would then be in violation of the fourth Guass-Markov Theorem condition (non-autocorrelation), indicating that OLS a suboptimal estimator relative to the models outlined in the next section of this report.

1.3 Regression Models with Autoregressive Components

Once the functional form of the spatial autocorrelation is identified, it is necessary to account for the appropriate type of autocorrelation exhibited in the data. Cliff and Ord (1981) caution analysts to determine whether the spatial process at play in the data is cause by "reaction or interaction" when choosing the appropriate spatial regression model. It is important to identify whether the spatial units are interacting with one another or they are reacting to some other variable not included in the model. The determination of what is driving spatial processes is often driven by both testing statistical hypotheses and the judgment of the researcher.

1.3.1 The Spatial Autoregressive Regression Model (SAR)

The spatial autoregressive regression (SAR) model is analogous to the Autoregressive Model (AR) in time series statistics.[4] In terms of spatial statistics, the spatial lag term, $\rho\mathbf{WY}$, is a weighted average of neighboring values (Anselin and Rey 2014). The SAR model in matrix notation is,

$$\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.4}$$

[4]The SAR model is sometimes referred to as the spatial autoregressive lag model.

with

$$-1 < \rho < 1 \tag{1.5}$$

where $\boldsymbol{W}$ is the n×n row-normalized spatial weights matrix, ρ is the scalar autoregressive lag parameter, $\mathbf{X}$ being the n×k matrix of explanatory variables, $\boldsymbol{\beta}$ is a k×1 vector of regression coefficient parameters, and $\boldsymbol{\epsilon}$ is a n×1 vector of independently and identically distributed residuals.

The spatial weights matrix, $\boldsymbol{W}$, is essentially a weighted contiguity matrix which relates observations of a variable at one point in space to other observations of that same variable at other points in space.[5] Matrix items, W_{ij}, will be zero whenever spatial units are not deemed neighbors.[6] In general, $\boldsymbol{W}$ will be large and populated with many zeros. Traditionally, the spatial weights matrix is row-standardized by dividing each row by its row sum.

When the scalar ρ=0, Equation 1.4 reduces to the standard linear regression equation. It is possible to estimate the the autocorrelation parameter, ρ with maximum likelihood estimation using the likelihood function (Griffith 1987). [7]

---

[5]In a contiguity matrix, neighboring spatial units are assigned a value of one (Geary 1954; Moran 1948).

[6]The definition of neighboring units depends on the type of contiguity used. Queen-based contiguity includes spatial units sharing both borders and vertices, while Rook-based contiguity only includes spatial units sharing borders.

[7]Griffith (1987, p. 30) notes that even when minimizing the log-likelihood function (Equation A.4), it is not possible to derive closed form parameter estimates for ρ and the asymptotic standard error of ρ. When estimating ρ, the partial derivative of the log-likelihood function does not reduce to a linear form, so it necessary to use non-linear optimization techniques on the log-likelihood function.

1.3.2 Spatial Autoregressive Error Regression Model (SAER)

When the regression error in one location is dependent on the error in another location, it is necessary to use the spatial autoregressive error regression (SAER) model. In the SAER model, the autocorrelated error term, $\boldsymbol{\mu}$ is a function of the autocorrelation parameter, λ, a matrix of spatial weights for paired observations, $\mathbf{W}$, the autocorrelated error term of another observation, and an identically and independently distributed error term, $\boldsymbol{\epsilon}$ (Odland 1988).

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu} \\ \boldsymbol{\mu} &= \lambda \boldsymbol{W} \boldsymbol{\mu} + \boldsymbol{\epsilon} \end{aligned} \tag{1.6}$$

with

$$\begin{aligned} \epsilon &\sim N(0, \sigma_\epsilon^2 I^2) \\ -1 < &\lambda < 1 \end{aligned} \tag{1.7}$$

To find a consistent estimate for $\boldsymbol{\beta}$ it is necessary to use spatially weighted least squares (sometimes referred to as spatial Cochrane-Orcutt) to estimate λ (Anselin and Rey 2014).

1.3.3 The General Spatial Process Regression Model

Both the SAR and the SAER models are special cases of the general spatial process regression model. The general spatial process regression model is essentially a spatial autoregressive model with a spatial autoregressive error term. Anselin (1988) specifies the general spatial process regression model as,

$$\mathbf{Y} = \rho \boldsymbol{W_1} Y + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}$$
$$\boldsymbol{\mu} = \lambda \boldsymbol{W_2} \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{1.8}$$

with

$$\epsilon \sim N(0, \sigma_\epsilon^2 I^2)$$
$$-1 < \rho, \lambda < 1 \tag{1.9}$$

where $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are n×n row-normalized spatial weights matrices, ρ is the scalar autoregressive lag parameter, λ is the scalar autoregressive error parameter, $\boldsymbol{\mu}$ is the autoregressive error term, $\mathbf{X}$ being the n×k matrix of explanatory variables, $\boldsymbol{\beta}$ is a k×1 vector of regression coefficient parameters, $\boldsymbol{\epsilon}$ is a n×1 vector of independently and identically distributed residuals.

For many applications, including both the spatial lag parameter and spatial error parameter is not necessary. Depending on the results of the Lagrange Multiplier (LM) test (described in section 1.4.3), it may be more appropriate to use the Spatial Autoregressive Regression Model (SAR) or the Spatial Autoregressive Error Model (SAER).

1.3.4 Maximum Likelihood Estimation of Spatial Autoregressive Regression Models

Due to the non-spherical variance-covariance matrix, OLS becomes inefficient. Fortunately, it is fairly straightforward to proceed using maximum likelihood estimation. Maximum likelihood estimates are asymptotically efficient, achieving the Cramér-Rao lower variance bound. Table 1 presents the log-likelihood functions we aim to maximize to derive parameter estimates.[8]

[Table 1 about here.]

There are several operational issues in the implementation of maximum likelihood estimation of spatially weighted least squares. In particular, the maximum likelihood estimate of β is conditional on the value of the scalar ρ or λ. As outlined by Anselin (1988), the procedure for estimating parameter values is slightly different for the SAR model relative to the SAER model.

1.3.4.1 Maximum Likelihood Estimation of the SAR model

For the SAR model, the operational issue stems from the first order maximum likelihood estimate of β being a function of ρ.

$$\beta = (X'X)^{-1}X'y - \rho(X'X)^{-1}X'Wy \tag{1.10}$$

Anselin (1988) summarizes the estimation procedure in five parts:

1. Carry out OLS of X on y to generate an estimate of β_o
2. Carry out OLS of X on Wy to generate an estimate of β_L
3. Use β_o and β_L to generate estimates of the residuals ϵ_0 and ϵ_L

[8] Maximizing a log-likelihood function is equivalent to maximizing the likelihood function. Typically, the log-likelihood function is easier than the likelihood function to manipulate algebraically.

4. Maximize the likelihood function with respect to ρ

$$\mathcal{L} = \alpha - (N/2)ln[(1/N)(\epsilon_0 - \rho\epsilon_L)'(\epsilon_0 - \rho\epsilon_L)] + ln|I - \rho W| \quad (1.11)$$

5. Given the estimate of $\hat{\rho}$, compute:

$$\begin{aligned} \beta &= \beta_0 - \hat{\rho}\beta_L \\ &and \\ \sigma^2 &= (1/N)(\epsilon_0 - \hat{\rho}\epsilon_L)'(\epsilon_0 - \hat{\rho}\epsilon_L) \end{aligned} \quad (1.12)$$

1.3.4.2 Maximum Likelihood Estimation of the SAER model

For the SAER model, the operational issue stems from the first order maximum likelihood estimate of β being a function of λ. After estimating λ, it is possible to estimate β using feasible generalized least squares (FGLS). Anselin (1988) outlines a seven step the procedure for generating $\hat{\beta}_{FGLS}$

1. Estimate $\hat{\beta}_{OLS}$ via OLS regression

2. Estimate $\hat{\epsilon}_{OLS} = y - X\hat{\beta}_{OLS}$

3. Maximize the likelihood function with respect to λ

$$\mathcal{L} = \alpha - (N/2)ln[(1/N)(\epsilon'_{OLS}(I - \lambda W)'(I - \lambda W)\epsilon_{OLS}] + ln|I - \lambda W| \quad (1.13)$$

4. Given $\hat{\lambda}$, carry out feasible generalized least squares to obtain $\hat{\beta}_{FGLS}$ where

$$\hat{\beta}_{FGLS} = [X'(I - \hat{\lambda}W)'(I - \hat{\lambda}W)]^{-1}X'(I - \hat{\lambda}W)'(I - \hat{\lambda}W)y \quad (1.14)$$

5. Compute $\epsilon_{FGLS} = y - X\hat{\beta}_{FGLS}$

6. Check for convergence

7. Given ϵ_{FGLS} and $\hat{\lambda}$ compute $\sigma^2 = (1/N)\epsilon'_{FGLS}(I - \hat{\lambda}W)'(I - \hat{\lambda}W)'\epsilon_{FGLS}$

1.4 Statistics for Quantifying Spatial Dependencies

A statistically significant result would imply that we can reject the null of randomness and independence of observations. It would then be appropriate to consider the dependent variable as being systematically organized across space. In other words, the pattern in the dependent variable observed is unlikely to have occurred if it was truly randomly distributed across space. Global spatial statistics estimate the degree to which the dataset is spatially organized in clusters of like-values. The most common statistic for testing spatial autocorrelation in continuous data is Morans I .[9]

1.4.1 Moran's *I*

Moran's I is a correlation coefficient between observations which are nearest neighbors (Moran 1950).

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.15}$$

[9]It should be noted that Morans I is one of many statistics measuring spatial dependency.

Morans I is asymptotically normal with an expected value of $-\frac{1}{(n-1)}$ under the null hypothesis of independently distributed observations (Griffith 1988; Cliff and Ord 1981; Moran 1950; Moran 1948). The theoretical sampling distribution can then be used to generate a confidence interval and a Z-statistic to test whether it is appropriate to reject the null hypothesis of independence of observations. Statistical significance of Morans I can also be determined through non-parametric means. Exact p-values of Morans I can be obtained from a random permutation test with a permutation distribution composed of $n!$ Morans I statistics.[10]

1.4.2 Local Moran's *I* (LISA)

In contrast to the global Moran's I outlined in section 1.4.1, there is a local indicator of spatial association (LISA) also known as the Local Moran's I.

$$I_i = y_i \sum_j W_{ij} y_j \tag{1.16}$$

This diagnostic measures the contribution of each spatial unit to the global Morans I (Equation 1.15). The sum of LISAs for all spatial units is proportional to the global Moran's I (Anselin 1995). Substantively, LISA helps the researcher identify clustering patterns in spatial data. Statistically significant values indicate clusters of *high-high*, *low-low*, and *high-low* pairs of spatial units.

[10] Due to large number of permutations necessary for most geographic analyses, it is prudent to use a Monte Carlo approach to approximating the permutation distribution (Anselin 2009).

The non-parametric test for statistical significance developed by Anselin (1995) tests the null hypothesis of no spatial autocorrelation between neighboring pairs.[11]

1.4.3 Lagrange Multiplier Test for Spatial Dependence

When determining the appropriate model to use for analysis, Anselin et al. (1996) suggest using the Lagrange multiplier (LM) test for spatial dependence. The LM test is used for testing hypotheses about parameters in the likelihood framework. More precisely, the LM test tests the hypothesis of a simpler model by maximizing the log likelihood subject to restrictions. When the LM statistic is large, the null hypothesis of a simpler model should be rejected. Compared to the Wald test and the Likelihood-ratio test, the LM test is the least stringent and most appropriate for testing model specifications (Engle 1980).

Anselin et al. (1996) note that the is no theoretical basis for the assumption of $W_1 \neq W_2$ in applied research. Since there is no substantial reason for this assumption in terms of this analysis, the LM statistic used for hypothesis testing will be simplified such that $W_1 = W_2 = W$. [12]

Model selection in this paper will rely on two LM tests presented in

[11]The test for statistical significant developed by Anselin (1995) takes a conditional randomization approach using randomized permutations.

[12]The only caveat to making this simplification is that the null of simultaneously testing for ρ and λ (the general spatial process regression model) cannot be tested using the LM test due to identification issues (Anselin et al. 1996).

Anselin et al. (1996). Equation 1.17 tests the hypothesis of a spatial lag ($H_o : \rho = 0$) in the presence of spatial disturbances. Equation 1.18 tests the hypothesis of spatial disturbances ($H_o : \lambda = 0$) in the presence of a spatial lag. Both tests allow for the parameter not of interest to be unrestricted.

$$LM_\rho = \frac{[\frac{\tilde{\mu}'Wy}{\tilde{\sigma^2}} - \frac{\tilde{\mu}'W\tilde{\mu}}{\tilde{\sigma^2}}]^2}{N\tilde{J}_{\rho\cdot\beta} - T} \tag{1.17}$$

$$LM_\lambda = \frac{[\frac{\tilde{\mu}'W\tilde{\mu}}{\tilde{\sigma^2}} - T(N\tilde{J}_{\rho\cdot\beta})^{-1}\frac{\tilde{\mu}'W\tilde{\mu}}{\tilde{\sigma^2}}]^2}{T[1 - T(N\tilde{J}_{\rho\cdot\beta})]^{-1}} \tag{1.18}$$

A more detailed explanation of the LM statistic used in this analysis is available in the Appendix.

In practice, there is a two-step procedure for implementing LM tests for spatial dependence. Depending on the results of the typical LM for spatial autocorrelation in the dependent variable or the error term, it may be necessary to use a more robust LM test.

First, it is necessary to test the null hypothesis of a linear regression model which does not account for autocorrelation relative to the SAR model and the SAER model with the LM test in Equations 1.17 and 1.18. If we fail to reject the null hypotheses in both tests, it is appropriate to use OLS. If we fail to reject the null hypothesis in one of the tests, but reject the null in the other test, then we should proceed in our analysis with the model that rejected the null.

If the null was rejected in both tests, it is appropriate to use a more robust form of the LM test for identifying the spatial process.[13] The results of the robust LM test should identify whether the spatial process generating autocorrelation is in the dependent variable or the error term. If there are two distinct spatial weight matrices specified for the autogressive lag term ($\rho \boldsymbol{W_1} y$) and the autoregressive error term ($\lambda \boldsymbol{W_2} \mu$), then it is possible to test the fit of the general spatial process regression model using the LM test for spatial dependence.

[13]The robust LM test for spatial dependence can be implemented in R using the package "spdep".

Chapter 2

Case Study: Turnout in Texas

Despite the obvious relevance of the election experience for subsequent democratic participation, the voting and elections literature has overwhelmingly focused on voting behavior rather than the actual act of voting itself. Since the controversial 2000 presidential election, there has been an increasing demand for information about improving the conduct of American elections. With only a decade-and-a-half of sustained attention by political scientists, our understanding of election administration has grown greatly. Most notably, research has focused on politically salient issues like turnout, residual vote rates, voter identification, and voter suppression. Although these issues are important and contemporaneous, persistent less visible problems plague the system and attract scant scholarly attention. It takes major election mishaps to garner attention to issues that have been of utmost concern to local election officials all along, such as creating foolproof ballots and machinery accessible to voters with disabilities.

2.1 Election Administration in the 21st Century

The decade of 2000-2010 saw unprecedented election reform (Montjoy 2010). One of the most substantial was the Help America Vote Act of 2002 (HAVA; P.L. 107-252). In a nutshell, "HAVA created the Election Assistance Commission (EAC), established a set of election administration requirements, and provided federal funding, but did not supplant state and local control over election administration (*The Help America Vote Act and Election Administration* 2015). As a result, the election administrative system has become increasingly complex, leaving the burden on local election administrators to navigate and implement changes (Montjoy, 2008).

In a survey of local election officials, Kimball et al. (2013, p.567) report that election officials interpret their policy environment as administratively burdensome due to "an ongoing set of unfamiliar requirements that have made their life more difficult. Currently, election administrators feel increased pressure to find quick-fixes and act as problem-solvers. Many administrators find it difficult to keep up with financial and labor costs associated with the requirements set by HAVA. Recent research shows that there are many non-trivial additional costs (new ballot forms, additional hours worked, rental space, etc.) associated with upgrading voting equipment. Furthermore, these additional costs vary year to year based on market prices. These rising costs and decreasing budgets are of utmost concern in an era where administrators have to oversee elections where partisan suspicions are high. As a result, the administrators have become handicapped in their ability to provide high quality elections. Like a straw that breaks the camel's back, these less visible but persistent problems can become catastrophic.

Underlying the concerns of local election officials about the state of their voting equipment is the notion that voting technology has a substantial impact on the quality of American elections. Voting equipment can quite literally be considered the machinery of democracy. Therefore it is important to understand the relationship between quantity and quality of voting equipment on voting behavior.

This paper will examine the impact of differences in resource allocation on turnout. I will use the term "resource allocation loosely to describe financial resources spent and labor mobilized to maintain or improve a jurisdictions voting equipment. Now that HAVA funding is no longer provided to localities at the intended capacity, there is substantial variability in the amount of resources available to localities to spend on voting equipment. In other words, some localities are resorting to austerity while others are free to make necessary purchases on a regular basis.

2.2 Data and Variables

I intend to examine the impact of the cost of voting equipment as well as the impact of purchasing vendor services on turnout and the Election Day experience. When localities purchase vendor services, it is for the most part optional. These services are purchased with the intent to improve efficiency and accuracy in conducting elections. There is, however, no academic research on whether these services are actually producing better experiences.

A unique dataset was created to measure the cost of voting equipment using data from local-level contracts for the acquisition of voting equipment. An open records request, also known colloquially as a Freedom of Information Act (FOIA) request, is the process by which a citizen may ask to obtain a copy or inspect documents that are considered to be public information, but are not made publicly available.[1] Municipal contracts are considered public information (the bidding process, however, is not). County and municipal level contracts for the acquisition of voting equipment will provide data on:

1. Cost per voting equipment unit
2. Geographic variability in cost
3. Number of units currently in use

[1] Although both terms are identical in terms of the type of request, the Freedom of Information Act is a federal law. Governmental transparency laws are referred to by different names depending on the state. For example, in Texas the Texas Public Information Act governs open records requests made to the state and local governments.

4. Services subcontracted to vendors

Variables that can be derived from these contracts include age of voting equipment, number of units per polling place, number of units per registered voter, dollars spent per registered voter, and dollars spent per capita on voting equipment.

Data collected from the contracts, was then merged with turnout data from the Secretary of State of Texas and demographic data from the United States Census Bureau. This is possible through the inclusion of geodesic place codes like Federal Information Processing Standards (FIPS) code.

2.2.1 Turnout in Texas

Due to the nature of federalism, where regional subunits of government jointly share authority with the national government, location in the United States determines a great deal about the manner in which citizens are represented. This is consistent with the Madisonian conception of American popular sovereignty in that power is distributed throughout the system. Moreover, Ewald (2009, p.97) argues, "uniformity is actually not a central value of American elections. Even during the Founding, the notion of a decentralized election system was not a controversial topic. Given this, it would be naive to consider measures gauging electoral participation, such as turnout, as taking on uniform values across the country. Figure 1 depicts the distribution of turnout across Texas counties. From the pattern observed, it is fair to consider turnout as a geographic phenomenon.

[Figure 1 about here.]

The traditional way of calculating turnout is by dividing the total number of voters by the voting age population (Burden and Neiheisel 2013). For the purposes of this study, turnout is calculated by dividing the total number of voters registered on Election Day. Since this study is primarily focused on measuring the impact of election administration variables on facilitating voters in casting their ballots, it is not necessary to include those who are not eligible to vote or do not wish to participate.[2] Turnout and registration data at the county level were downloaded from the Texas Secretary of State's Elections Division.

We should expect to find "neighborhood effects" in county-level turnout for two reasons in particular:

1. Not all electoral districts are nested within county boundaries.

2. County boundaries were not constructed (nor redefined) to accommodate socio-political communities.

Figure 2 is a map of the City of Austin, TX. Highly competitive races for Austin city government will drive turnout not only in Travis County (which contains the majority of the city), but also in Bastrop, Hayes, and Williamson counties (which also contain a some of the city). As a result, this scenario

[2]Since voters must be registered well in advance of Election Day, the exact number of registered voters known on Election Day.

would lead to a clustering pattern in turnout across the spatial units, y_{Travis} and $y_{Williamson}$.

[Figure 2 about here.]

2.2.2 Independent Variables

Investment in voting equipment per registered voter. Voting equipment is considered any piece of hardware used to facilitate the counting and casting of ballots.[3] The total investment in voting equipment is the dollar amount spent on the hardware of most current system of voting equipment in use in 2016 purchased by counties from voting equipment vendors.

Vendor. A dummy variable for voting equipment vendor (analogous to manufacturer) is included in the model to control of vendor specific effects. To date, there are no published studies in any political science journal regarding the impact of vendors on any aspect of the electoral process. Election equipment in the United States is almost exclusively purchased from private-sector vendors. When a jurisdiction purchases voting equipment, they are actually purchasing the hardware and software along with a variety of services for the initial implementation and long-term service and support of the system. In

[3]Software was not included in the calculation of investment in voting equipment due to the disparities in licensing agreements across vendors. For instance, some vendors offer perpetual licenses, while others may only provide annual licenses. In addition, in the FOIA request made to counties for voting equipment contracts did not make an explicit request to obtain copies of contracts for software licenses. For many counties, this additional request would have complicated the data gathering task substantially.

other words, not only is voting equipment purchased, but so are services provided by the vendors to maintain the equipment. Unlike other industries, customers cannot "substitute away from voting equipment when vendors increase their prices. Because voting equipment uses proprietary software, local election officials also cannot mix and match products from different companies. Therefore, firms with large product catalogues are desirable.

In the state of Texas, there are three main vendors:

1. Hart Intercivic, Inc.("Hart")
2. Election Systems & Software, LLC.(ES&S)
3. Dominion Voting Systems, Inc. ("Dominion") [4]

The coding scheme for vendors does not, however, include Dominion due to issues with multicollinearity. As only a small minority of counties use Dominion hardware, it is unlikely that the model will be able to identify the influence of Dominion without more data. Instead, the variable "Hart" is a dummy variable coded "1" for counties using Hart voting systems, while counties using both ES&S and Dominion voting systems are coded as "0". The choice to code for Hart voting systems rather than ES&S is purely discretionary.

Vendor Services. Depending on the vendor, counties may elect to purchase additional election services to facilitate the planning and conduct of

[4]Dominion Voting Systems purchased Premier Election Solutions, formerly Diebold Election Systems, Inc. and Sequoia Voting Systems, Inc. in 2010

elections. Across most vendors, services for training, Election Day support, ballot production, and project management, are available for purchase. To date, there is no academic research examining the impact of vendor services on turnout.

Mode of counting and casting of ballots. There are four possible manners in which ballots may be cast and counted in Texas:

1. Direct-recording electronic voting machine (DRE)
2. Paper-based system using Optical Scanners
3. Both DRE and paper-based systems made available to voters (Mixed)
4. Hand-counted paper ballots

The model includes a sole dummy variable labeled "Mixed" to indicate counties using both DRE and paper-based systems. The choice to include only one indicator in the model is due to identification issues with vendor offerings in voting equipment.[5]

Age of voting system. The age of a voting system is determined from the date on the first county contract for the acquisition of voting equipment of the model currently in use in 2016. The start of a voting system's lifespan would commence following the first June (the start of the annual election

[5]The vendor Hart does not offer a paper-based system with ADA compliant ballot marking devices.

cycle) the county was in possession of the equipment. For instance, if a county purchased equipment in September of 2006, then the equipment was coded as nine years old in 2016. It should be noted that many counties made subsequent minor purchases to supplement and replace devices to meet state and federal guidelines.

Precincts per registered voter. The building block of all electoral districts is the precinct. In every precinct, all voters receive the same ballot. Alternatively, all voters in a precinct vote for all the same offices. The number of precincts is divided by the number of registered voters as a manner of standardization across counties.

Average number of machines per precinct. The total number of voting machines is the sum of either all DRE machines or Ballot Marking Devices depending on the county's chosen mode of ADA compliance.[6] The total number of voting machines was provided by the Annual Voting Systems Report published by the Secretary of State of Texas. The total number of machines is then divided by the total number of precincts in each county.

Demographic variables. In addition to variables measuring elections, demographic controls are included in the model. County level demographic values was obtained from the 2016 American Community Survey 5-year estimates data compiled by the United States Census Bureau.

[6]HAVA requires all counties to have ADA compliant voting equipment available to voters with disabilities.

Chapter 3

Results

3.1 Ordinary Least Squares Regression

The aim of this study is to evaluate the impact of resource allocation by election administrators on county turnout. The results of an OLS regression on turnout is presented in Table 2. Both the standardized and unstandardized coefficient estimates are reported. Although preliminary, there is statistically significant evidence in support of a positive association of investment in voting equipment per registered voter and county-level turnout. In terms of demographic control variables commonly associated with turnout, there is a statistically significant and positive relationship between the percentage of whites and per capita income and turnout.

[Table 2 about here.]

3.2 Moran's I Statistics

Table 3 presents the results of a two-sided Moran's I test. The weights matrix uses queen contiguity which is when the weighting scheme includes all neighbors sharing at least one border and all neighbors sharing at least one vertex.[1] The results indicate that there is indeed spatial autocorrelation present in the data. The results are nearly identical across approaches. The positive statistics across Moran's I specifications and deviation from the expected value indicate that there are instances of clustering of *high-high* and *low-low* values among neighboring units. Given this, the results of Table 2 are likely to be inefficient. It is therefore prudent to proceed with an alternative approach to parameter estimation other than OLS. The LM test presented in Section 3.4 will determine the appropriate spatial autoregressive regression model to use on the data.

[Table 3 about here.]

3.3 Local Moran's *I*

Figure 3 depicts the LISAs for turnout in counties in Texas. Statistically significant LISAs are depicted by a solid fill. "Cold spots" are indicated by solid blue fill and "hot spots" are indicated by solid red fill. The pattern presented suggests that there are three "cold spots" (*low-low*) and four "hot spots" (*high-high*). In general, there appears to be higher turnout in the northern and central regions of Texas, while the western and southern regions of Texas appear to have lower turnout.

[Figure 3 about here.]

[1]This is in contrast to Rook contiguity where only neighbors sharing borders are included in the weighting scheme.

3.4 Lagrange Multiplier Test

Table 4 presents the results of the LM test for spatial dependence for both model specifications including spatial lag and spatial error parameters. The LM statistics under the null hypothesis follow a χ_1^2 distribution. Under the LM test specified in Equations 1.17 and 1.18, the null hypothesis was rejected for both tests. It is therefore prudent to proceed with the robust LM test for spatial dependence. The results of the LM test for autoregressive error rejects the null hypothesis of the OLS regression model (no autoregressive error term) in favor of the SAER model. Given that the robust LM test for the SAR model failed to reject the null hypothesis, this analysis will proceed in Section 3.5 with the SAER model.

[Table 4 about here.]

3.5 Spatial Autoregressive Error Regression

Table 5 presents the results of the SAER model on 2016 turnout in Texas. Compared to the AIC for the OLS regression model in Table 2, the AIC for the SAER model in Table 5 is lower, indicating that accounting for spatial autocorrelation in the error term improves the fit of the model. The spatial dependence coefficient, λ is positive and statistically significant. This result is consistent with the value of the Moran' I statistic reported in Table 3.

[Table 5 about here.]

After accounting for spatial autocorrelation, the variable "Total population" reaches statistical significance, thus producing different results from those based on the OLS specification in Table 2.

These results further support the hypothesis that increasing resources in election administration can have a positive impact on turnout. The coefficient estimates of "Investment in Equip. / Reg. Voters" is positive and statistically significant. A one standard deviation in investment in election equipment per registered voter translates into a 1.95% in county-level turnout.

Chapter 4

Conclusions

Using county-level turnout data from Texas, the results presented in Table 5 suggest a positive relationship between spending on voting equipment and turnout. Substantively, these results are non-trivial. Counties that invest more resources into elections appear to have higher levels of turnout than those who invest less, all else being equal. It should be noted that these results suggest a causal relationship. Unfortunately, this results cannot confirm this conclusion. Analogous to the age old question of the *chicken or the egg*, this analysis cannot detect whether increasing resources invested in elections is the definitive causal mechanism explaining increases turnout.

Methodologically, the results of this study suggest that modeling spatial data appropriately does have substantive implications on regression analysis. By including specifications for spatial autocorrelation in the error term the fit of the model was improved.

Considering the recent deluge of publicly accessible *big data* produced by governmental entities, it is imperative for researchers to understand how to recognize and model spatial data. Political scientists studying local government, in particular, should be cognizant of the difficulties in dealing with spatial data. As there are roughly 3,000 counties in the United States, there exists the possibility of a complex scheme of spatial dependencies that must be taken into account in any county-level analysis.

Tables and Figures

Figure 1: Turnout in Texas in 2016

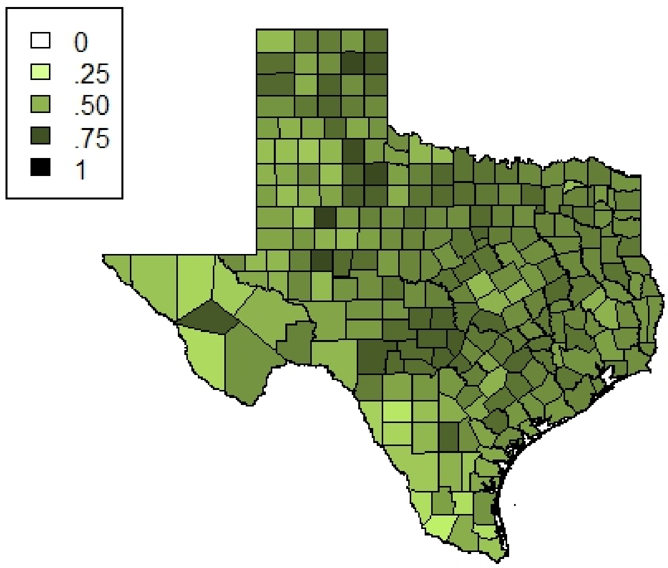


Figure 2: Boundary Map of Austin, TX

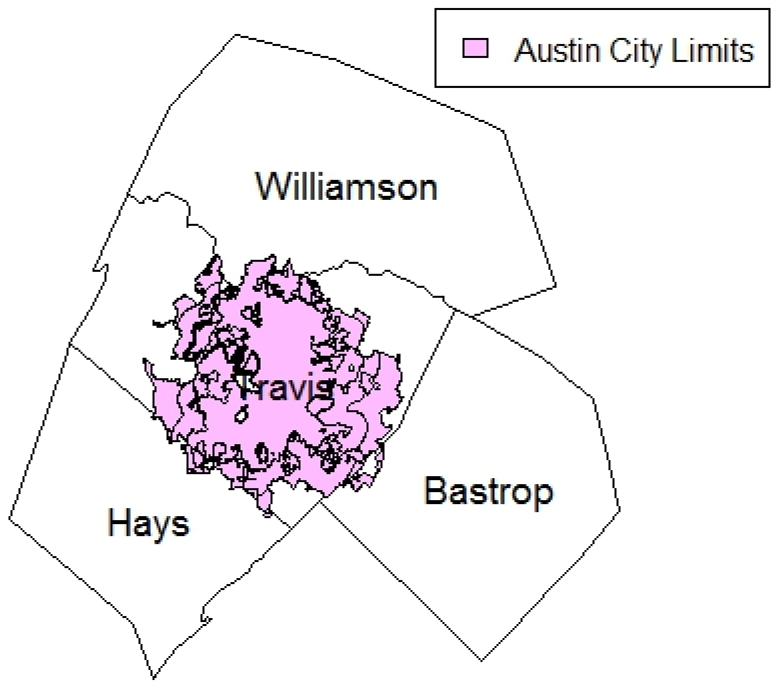
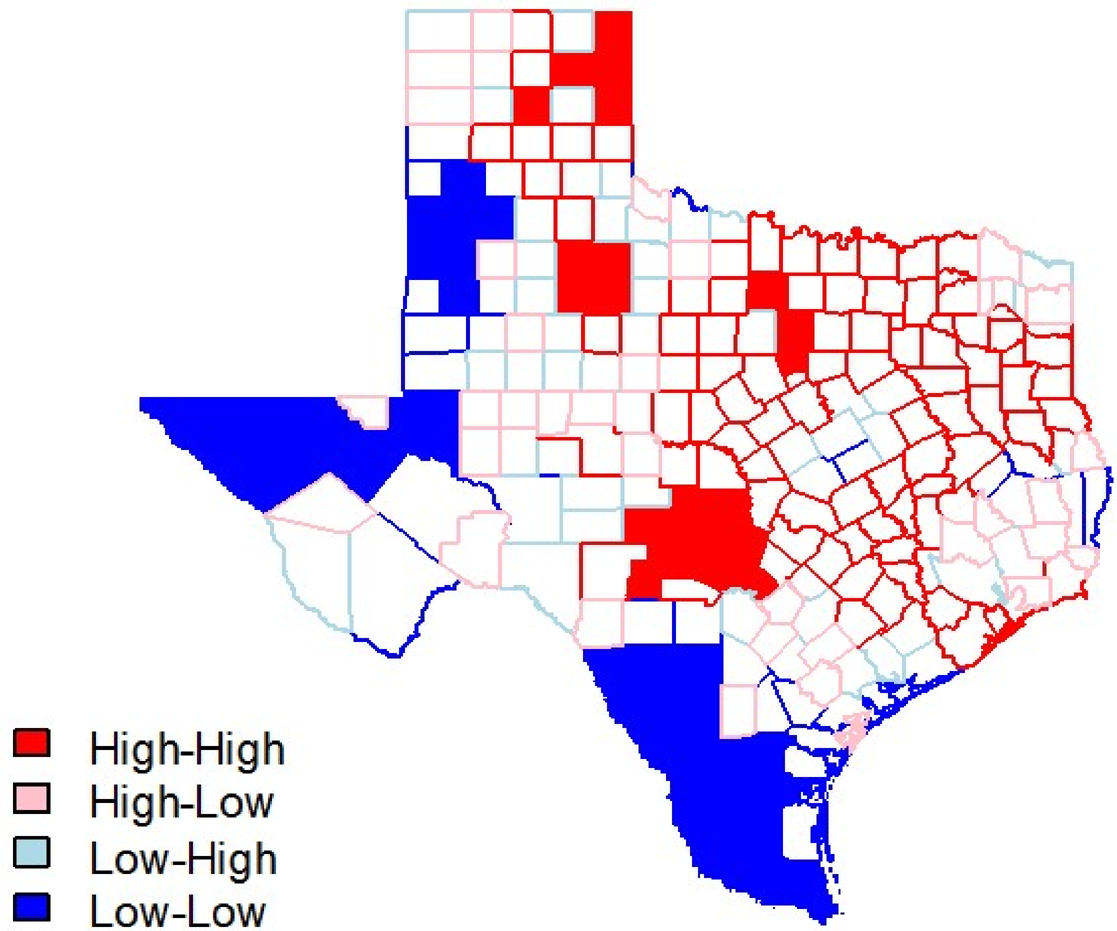


Figure 3: Local Moran's I for Turnout in Texas



Note: Counties with statistically significant Local Moran's I values are indicated by a solid color fill.

Table 1: Log-Likelihood Functions for Spatial Autoregressive Regression Models

Model	Log-Likelihood Function
General	$-\frac{n}{2}ln(2\pi) - \frac{1}{2}ln(\sigma^2) + ln(I-\rho W_1) + ln(I-\lambda W_2) - \frac{1}{2\sigma^2}[(I-\rho W_1)y - X\beta]'(I-\lambda W_2)'(I-\lambda W_2)[(I-\rho W_1)y - X\beta]$
SAR	$-\frac{n}{2}ln(2\pi) - \frac{1}{2}ln(\sigma^2) + ln(I-\rho W_1) - \frac{1}{2\sigma^2}[(I-\rho W_1)y - X\beta]'[(I-\rho W_1)y - X\beta]$
SEM	$-\frac{n}{2}ln(2\pi) - \frac{1}{2}ln(\sigma^2) + ln(I-\lambda W_2) - \frac{1}{2\sigma^2}[y - X\beta]'(I-\lambda W_2)'(I-\lambda W_2)'(y - X\beta)$

Table 2: Ordinary Least Squares Regression of Turnout in Texas in the 2016 General Election

	Unstandardized		Standardized	
	B	Std. Error	B	P-value
Intercept	-4.433	15.15	61.12	0.770
Election Administration:				
Machines/ Registered Voters	-35.95	28.82	-0.390	0.214
Registered Voters / Precincts	0.000	0.000	29.569	0.284
Hart	-0.123	0.788	-0.061	0.876
Mixed System (DRE + paper)	-0.507	0.772	-0.197	0.512
Investment in Equip. / Reg. Voters	0.105	0.019	1.816	0.000***
Age of System	0.734	0.384	0.582	0.057
Vendor Services:				
Training	0.864	1.078	0.864	0.424
Election Day Support	-0.594	0.772	-0.594	0.443
Voter Outreach	-0.920	0.865	-0.920	0.289
Project Management	0.035	0.806	0.035	0.965
Demographics:				
Total population	0.000	0.000	2.622	0.148
Median Age	0.102	0.064	0.608	0.113
% College	0.000	0.000	-1.308	0.468
% White	0.433	0.155	9.138	0.006**
% Black	0.311	0.165	2.076	0.060
% Hispanic	0.265	0.152	6.122	0.083
Per Capita Income	0.001	0.000	2.964	0.000***
R-squared: 0.6623				
AIC: 1428.495				
N=245 (9 deleted due to missing data)				

Table 3: Moran's I statistics

	Moran's I Statistic	$E(I)$	Deviation from $E(I)$	P-value
Parametric approach:				
Dependent variable	0.326	-0.004	0.330	0.001**
Error term	0.159	-0.017	0.176	0.000***
Monte Carlo approach:				
Dependent variable	0.326	-	-	0.001***
Error term	0.159	-	-	0.001***

Table 4: Lagrange Multiplier Test for Spatial Dependence

Model	Lagrange Multiplier Test
Autoregressive Lag	10.643***
Autoregressive Error	16.25***
Robust LM test:	
Autoregressive Lag	0.657
Autoregressive Error	6.264*

Table 5: Spatial Simultaneous Autoregressive Error Regression of Turnout in Texas in the 2016 General Election

	Unstandardized		Standardized	
	B	Std. Error	B	P-value
Intercept	3.749	14.436	61.210	0.795
Election Administration:				
Machines/ Registered Voters	-27.278	25.554	-0.296	0.286
Registered Voters / Precincts	0.000	0.000	33.675	0.175
Hart	0.136	0.715	0.067	0.849
Mixed System (DRE + paper)	-0.624	0.718	-0.243	0.385
Investment in Equip. / Reg. Voters	0.113	0.019	1.950	0.000***
Age of System	0.611	0.350	0.484	0.081
Vendor Services:				
Training	0.964	1.005	0.964	0.338
Election Day Support	-0.295	0.700	-0.295	0.673
Voter Outreach	-1.322	0.791	-1.322	0.095
Project Management	-0.033	0.756	-0.033	0.965
Demographics:				
Total population	0.000	0.000	3.277	0.048*
Median Age	0.079	0.065	0.470	0.224
% College	0.000	0.000	-2.161	0.188
% White	0.364	0.146	7.682	0.013*
% Black	0.256	0.158	1.710	0.106
% Hispanic	0.197	0.197	4.551	0.171
Per Capita Income	0.001	0.000	2.988	0.000***
Lambda	0.377	0.087	0.377	0.000***

Nagelkerke pseudo-R-squared: 0.70534
AIC: 1414.473
N=245 (9 deleted due to missing data)

Appendices

Appendix A

Maximum Likelihood Estimation

A.1 Probability density function of Multivariate Normal

$$f_x(x_1, ..., x_k) = \frac{exp(-\frac{1}{2}(x_k - \mu)\Sigma^{-1}(x_k - \mu)}{\sqrt{(2\pi)^k|\Sigma|}} \tag{A.1}$$

Where

$$|\Sigma| = det(\Sigma) \tag{A.2}$$

A.2 Likelihood function for Multivariate Normal (0, Σ)

$$\begin{aligned} \mathcal{L}(0, \Sigma|X) &= \prod_{k=1}^{n} f(x_k|0, \Sigma) \\ &= (2\pi)^{-\frac{nk}{2}}|\Sigma|^{-\frac{n}{2}}exp[-\frac{1}{2}\sum_{k=1}^{n} x_k\Sigma^{-1}x_k] \end{aligned} \tag{A.3}$$

A.3 Log-likelihood function for Multivariate Normal (0, Σ)

Maximizing the likelihood function is equivalent to minimizing the log-likelihood function.

$$\begin{aligned} ln\mathcal{L}(0, \Sigma|X) &= n[-\frac{k}{2}ln(2\pi) - ln(|\Sigma|) - \frac{1}{2}\sum_{k=1}^{n} x_k \Sigma^{-1} x_k] \\ &\propto -\frac{k}{2}ln(2\pi) - ln(|\Sigma|) - \frac{1}{2}\sum_{k=1}^{n} x_k \Sigma^{-1} x_k \end{aligned} \tag{A.4}$$

Appendix B

Lagrange Multiplier Test

The general approach to the Lagrange Multiplier test (Anselin et al. 1996).

Null hypotheses:

$$\begin{aligned} H_o :&\theta_o = (\beta', 0, 0) \\ &\theta_{Lag} = (\beta', 0, \rho) \\ &\theta_{Error} = (\beta', \lambda, 0) \end{aligned} \tag{B.1}$$

Where the matrix θ is the parameter vector, β is the vector of regression coefficients, the scalar λ is the spatial disturbance coefficient, and the scalar ρ is the the spatial lag coefficient. The derivative of the likelihood function for θ:

$$\delta(\theta) = \frac{\delta \mathcal{L}(\theta)}{\delta \theta} = \begin{bmatrix} \frac{\delta \mathcal{L}(\theta)}{\delta \beta} \\ \frac{\delta \mathcal{L}(\theta)}{\delta \lambda} \\ \frac{\delta \mathcal{L}(\theta)}{\delta \rho} \end{bmatrix} \tag{B.2}$$

The Jacobian matrix for θ:

$$J(\theta) = -E[\frac{1}{N} \frac{\delta^2 \mathcal{L}(\theta)}{\delta \theta \delta \theta}] = \begin{bmatrix} J_\beta & J_{\beta\lambda} & J_{\beta\rho} \\ J_{\lambda\beta} & J_\lambda & J_{\lambda\rho} \\ J_{\rho\beta} & J_{\rho\lambda} & J_\rho \end{bmatrix} \tag{B.3}$$

Bibliography

Anselin, L., and Sergio Joseph Rey. 2014. *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. Chicago, IL: GeoDa Press LLC.

Anselin, Luc. 1988a. “Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity”. *Geographical Analysis* 20 (1): 1–16.

—. 1995. “Local Indicators of Spatial Association - LISA”. *Geographical Analysis* 27 (2): 93–115.

—. 1988b. *Spatial econometrics: methods and models*. Boston, MA: Kluwer Academic Publishers.

—. 2009. “Spatial Regression”. In Fotheringham and Rogerson 2009, 255–275.

Anselin, Luc, et al. 1996. “Simple diagnostic tests for spatial dependence”. *Regional Science and Urban Economics* 26 (1996): 77–104.

Berry, William D. 1993. *Understanding Regression Assumptions*. Vol. 07-092. Newbury Park: Sage Publications.

Bivand, Roger, Jan Hauke, and Tomasz Kossowski. 2013. “Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods”. *Geographical Analysis* 45 (2): 150–179. `http://www.jstatsoft.org/v63/i18/`.

Bivand, Roger, and Gianfranco Piras. 2015. “Comparing Implementations of Estimation Methods for Spatial Econometrics”. *Journal of Statistical Software* 63 (18): 1–36. `https://www.jstatsoft.org/v63/i18/`.

Burden, Barry C., and Jacob R. Neiheisel. 2013. “Election Administration and the Pure Effect of Voter Registration on Turnout”. *Political Research Quarterly* 66 (1): 77–90.

Cliff, Andrew David, and J. K. Ord. 1981. *Spatial processes: models & applications*. London: Pion.

Engle, Robert F. 1980. “Hypothesis Testing in Spectral Regression; the Lagrange Multiplier Test as a Regression Diagnostic”. In Kmenta and Ramsey 1980, 309–321.

Ewald, Alec C. 2009. *The Way We Vote: The Local Dimension of American Suffrage*. Nashville, TN: Vanderbilt University Press.

Fotheringham, A. Stewart, and Peter Rogerson. 2009. *The SAGE Handbook of Spatial Analysis*. London: SAGE Publications, Ltd.

Geary, R. C. 1954. "The Contiguity Ratio and Statistical Mapping". *The Incorporated Statistician* 5 (3): 115–146.

Griffith, Daniel A. 1988. *Advanced spatial statistics: special topics in the exploration of quantitative spatial data series*. Boston, MA: Kluwer Academic Publishers.

— . 1987. *Spatial autocorrelation: a primer*. Washington, D.C: Association of American Geographers.

Hijmans, Robert J. 2017. *raster: Geographic Data Analysis and Modeling*. R package version 2.6-7. `https://CRAN.R-project.org/package=raster`.

Kimball, David C., et al. 2013. "Policy Views of Partisan Election Officials". *The UC Irvine Law Review* 3 (3): 551–574.

Kmenta, Jan. 1971. *Elements of Econometrics*. New York, NY: Macmillan Publishing Co.,Inc.

Kmenta, Jan, and James B. Ramsey. 1980. *Evaluation of Econometric Models*. Academic Press.

Montjoy, Robert S. 2010. "The Changing Nature and Costs of Election Administration". *Public Administration Review* 70 (6): 867–875.

— . 2008. "The Public Administration of Elections". *Public Administration Review* 68 (5): 788–799.

Moran, Patrick A.P. 1950. "Notes on Continuous Stochastic Phenomena". *Biometrika* 37 (1/2): 17–23.

— . 1948. "The Interpretation of Statistical Maps". *Journal of the Royal Statistical Society. Series B (Methodological)* 10 (2): 243–251.

Odland, John. 1988. *Spatial autocorrelation*. Newbury Park, CA: Sage Publications.

The Help America Vote Act and Election Administration. 2015. Congressional Research Service.